Inducing a Cline from Corpora of Political Manifestos

Sofie Van Gijsel* & Carl Vogel†

Abstract

Techniques from corpus linguistics are applied to the analysis of a number of European right-wing parties in an effort to extend methods for ranking parties on a left-right spectrum within and across countries and languages. Focus is placed on parties not in government, and analysis is derived from corpora derived from election manifestos published by those parties. The techniques applied are objective in that they apply statistical measures with confidence tests to objectively quantifiable linguistic features of the documents. Valid applicability of the techniques is demonstrated. The methods are then used to estimate pairwise similarity of a number of European political parties, including cross-national comparisons.

1 Introduction

We report on application of recent corpus linguistic methods to the analysis of a number of European right-wing party manifestos. In recent manifesto research aiming at estimating the policy positions of governmental parties of a nation in an objective way, computerized approaches have been used to locate parties on *a priori* established policy dimensions (see any number of articles in Laver, 2001a). We focus on the relatively unresearched policy space of (often) small, non-governmental, right-wing parties of a number of European countries. An aim is to identify objective means to rank these parties on a political spectrum using only limited data available from such parties. We use an inductive method which treats election manifestos as corpora to be analyzed. Thus, instead of in ideological terms, the manifestos are compared on the basis of *linguistically quantifiable features.* Clearly, ideological issues enter in the selection of parties whose manifestos are examined, but beyond these pre-theoretic choices, content-free statistical techniques are used to rank the level of similarity between the parties.

An initial methodological question is in determining whether it is legitimate to consider the right-wing manifestos, all clearly belonging to one subgenre, as 'corpora' which are distinguishable on the basis of significant linguistic differences. As recent research in corpus linguistics shows (Kilgarriff, 2001), in order to validly compare corpora, their *internal homogeneity* has to be larger than the distance between them. To measure the within-corpus distances, we apply recently proposed *authorship identification techniques* (AID), attempting to assign subparts of the manifestos correctly. This way, it is possible to cross-validate recent attributional research which shows that substrings of words are excellent author discriminators. We establish the internal homogeneity of the corpora, prerequisite for measuring the similarity levels among them.

A second question is then if a *corpus similarity measure* can be applied to evaluate the distance between the different parties, both on a national and a cross-national level. The recently proposed *Chi by Degrees of Freedom similarity measure* (Kilgarriff (2001); hereafter, $\frac{\chi}{d.f.}$) gives a ranking which we will attempt to interpret as an indication of the position of the different parties in a common policy space. The results suggest encouraging potential for new methods in analyzing manifestos in political science and other fields in which text-based induction of partially-ordered position-spaces is useful. We argue the objective analysis of small, 'real language' sets of texts as corpora, is an interesting, albeit challenging field of corpus linguistics.

*Quantitative Lexicology and Variational Linguistics, Katholieke Universiteit Leuven, Belgium: Sofie.VanGijsel@arts.kuleuven.ac.be

†Computational Linguistics Group & Centre for Computing and Language Studies, Trinity College, U. of Dublin: vogel@tcd.ie

2 Manifesto Research in Political Science

Manifesto analysis is considered a fruitful way of gaining insight into the positions of political parties in one policy space (Mair, 2001; Laver, 2001b). *The Manifesto Research Group* collects and analyzes political programs by way of *comparative content analysis*, classifying each 'quasi-sentence' according to a coding scheme of 56 categories, which belong to *a priori* established dimensions of the policy space (e.g. economical left-right, social liberal-conservative). The rationale behind this system is *salience theory* (Budge, 2001), which states that the *salience* of an issue in the manifesto provides information about the *position* of the party on that issue. Yet, this theory can be criticized: for example, immigration will be a hot issue for many parties—especially for the researched right-wing parties—but mentioning the issue in the program does not automatically point at being 'in favor' or 'against' it. This method also requires a large amount of human coding effort, which is time and money consuming, without being completely objective. Therefore, recent methods analyze the manifestos in a more quantitative way.

A first improvement is the *computerized content analysis* proposed by Laver and Garry (2000). On the basis of two reference texts or manifestos of parties for which the position on a number of pre-established policy dimensions is known *a priori*, the researchers make up a *keyword list*,[1] which will then be used to code other manifestos or 'virgin' texts. Yet, the composition of the keyword dictionary is not only time consuming, but also, the validity of the analysis is highly dependent on the keywords, which are sensitive to both the substantive and the temporal context of the reference manifestos.[2] Therefore, Laver, Benoit, and Garry (2003) recently have proposed a *probabilistic dictionary approach*, measuring the relative frequency of *all* the words in the references texts. For the analysis of 'virgin' texts, the policy position is then determined on the basis of the scores for all the words which are given a certain score on a dimension under investigation on the basis of the reference texts. This method allows rapid analysis and reanalysis of large quantities of texts. It is also applicable to non-English texts, an advantage if manifestos are compared cross-nationally. Yet, the reliability is still highly dependent on the choice of reference texts. Positioning virgin texts on *a priori* established dimensions, abstracted from reference texts, might be a good approach for well-researched policy spaces, but for the analysis of the often small and non-governmental right-wing parties analyzed in this project, this is not optimal.

Instead of using pre-established dimensions, we attempt to analyze the manifestos inductively into a partial-ordering, treating the complete texts as corpora (also sensitive to text choice, but because of the parties analyzed, this amounts to all of the available text, rather than choice), the distances among which can be measured. The distances can only be interpreted *a posteriori*.

3 Authorship Identification Techniques (AID)

As explained, the internal homogeneity of the manifestos has to be established before a valid corpus linguistic comparison is possible. We use a number of AID techniques to prove that the within-corpus distances are smaller than the those between the manifestos. First, a short overview and discussion of AID methods used in the analysis of style or *stylometry* will be given. Oakes (1998) and Holmes (1998) (for example) provide more comprehensive overviews. The methods we adopt are outlined in §3.2; later, §4 and §5 detail our analysis.

[1] Every word which occurs at least twice as many times in the right- or the left-wing reference text is classified as a right- or left-wing keyword respectively.

[2] See Van Gijsel (2002, p. 82-88) for the implementation of a keyword dictionary for Dutch, as devised by de Vries (1999). The results show that for the analysis of right-wing party manifestos of Belgium and Flanders (the Dutch-speaking part of Belgium), which entails a cross-national and temporal extension, the keyword dictionary does not give valid results.

3.1 Overview of the AID techniques

Stylometry as an AID technique dates at least to 1851, when the logician de Morgan suggested that the authenticity of some letters of St Paul might be tested comparing the *word length*. Yule (1944) developed a measure of vocabulary richness, K, based on the probability that any randomly selected pair of words are identical. Over the years, a number of other *vocabulary richness measures* as discriminators have been proposed, such as, for example, the *type-token ratio* or the proportion of *unique words* to the total size of the vocabulary used (e.g Morton, 1986), although more recent research (e.g. Holmes, 1998) shows that these techniques are not reliable, being highly dependent on the choice and length of the texts under analysis. Mosteller and Wallace (1964) famously attributed of the purposely anonymous, disputed *Federalist Papers* to Madison instead of Hamilton, on the basis of a *probabilistic* analysis of the most frequent words. These were mainly *function words*, which are rather unconscious and therefore effective markers of authorship. While most measures take the *lexical item* (or pre-terminal lexical categories as *parts-of-speech*) as the unit of analysis, some recent methods focus on *sublexical* units, especially *letter uni-and bigrams*. Without requiring syntactic or lexical analysis, these elements are easily and objectively quantifiable, while being useful for texts of varying and limited length (e.g. Forsyth (1997), Khmelev and Tweedie (2001), Chaski (1998)).

In literary stylistics, the *Cusum technique* (Farringdon, 1996) has been developed; it graphically plots the *average sentence length* of an author's sample, superimposed by plots for the frequency of a selected 'linguistic habits' of the author, such as the use of two and three letter words. The technique has been criticized for being labor intensive and highly subjective, e.g. with regard to the choice of the (limited) number of sentences analyzed, choice of selected linguistic habits and the interpretation of the plots (Canter, 1992; Chaski, 1998). Foster's (2001) analysis of the '*literary DNA*' of a writer is akin to the Cusum method and can similarly be criticized for being subjective and unscientific. Foster claims to uncover authorship on the basis of '*external'* (e.g. the historical background of a writer) and '*internal'* evidence (e.g. characteristics such as punctuation habits), but his recent incorrect attribution of 'A Funeral Elegy' to Shakespeare instead of to John Ford reveals the methodological unreliability of his method.

3.2 The AID techniques implemented

We have explored AID methods availing of letter unigrams and bigrams, since they could be applied cross-linguistically, and without subjective content based judgements, to small, unequally-sized texts. Thus, *letter unigrams* and *letter bigrams* were counted. Further, *word unigrams* were counted, to test if substrings give better results than word counts.

McCombe (2002) sought cross-validation of a number of AID techniques and confirmed recent work (e.g. Chaski, 1998) in that letter uni- and bigrams perform remarkably better than, in that order, word unigram frequency, syntactic tagging, higher n-grams or keywords as metric bases for predicting authorship of disputed texts. We used McCombe's software to test the validity of different AID methods in assigning arbitrarily selected subparts of the manifestos to the correct party. For detailed and user-oriented descriptions of its functionality see McCombe (2002) or Van Gijsel (2002). The program takes an input file consisting of names of plain text files, labeled to encode one or more uncontested categories, or as files to be categorized. Given input parameters (e.g. letter vs. word n-gram analysis, the value of n to n-gram, etc.), the texts are concordanced and frequency analyzed. The program's output is a pairwise ranking, giving the similarity of the various corpora in reverse magnitude, as calculated by $\frac{\chi}{d.f.}$.[3] Here, three rankings are given (letter uni- and bigrams and word unigrams), constituting a *rank list*.

[3]The $\frac{\chi}{d.f.}$ measure instead of simply χ^2 is used, since this takes into account both the χ^2 value and the frequency information of the corpora. This is useful for natural language corpora, like the manifestos, which are inherently non-randomly distributed (Kilgarriff & Salkie, 1996)

This rank list is the input for two statistical tests, which compare results of the tests as run with a range of parameter values. First, the *ratio* between the average of the similarity scores for all the pairs of corpora in the same uncontested category and the average similarity scores for all the pairs of corpora in distinct uncontested categories.[4] The larger the ratio, the more suggestive the measure is. McCombe (2002, p. 37) notices that the *ranking* of the assignment scores is often a more direct indication of the attributional result. A second test is the *Mann-Whitney test* (also called the *Wilcoxon rank sums test*; see Oakes, 1998),[5] which gives an overall *significance measure* for each of the three methods, while also outputting a more detailed list of significance measures for each of the three methods, showing the probability of the assignment of each of the anonymously coded texts to the different authors.

4 Analysis of the Manifestos Using AID Techniques

4.1 Data Collection

The manifestos were collected by downloading the texts from their respective party websites. To keep the human intervention to a minimum, the (thematic) subparts of the websites were kept intact as separate files of comparable size, but the number of themes by party varied. In this paper, the analysis of the Dutch language manifestos is addressed, both on a national and a cross-national level. For The Netherlands we analyzed the manifestos of the parties *Lijst Pim Fortuyn* (*List Pim Fortuyn, LPF*) and *Leefbaar Nederland (Liveable Netherlands, LN)*. The LPF-manifesto consists of a single text of a little under 4,000 words, while the LN-text contains 10 subparts (just over 10,000 words in total). The Belgian party manifesto of the *Vlaams Blok* (*Flemish Block, VB*) was downloaded in 13 chunks, amounting to more than 20,000 words.[6]

4.2 Analysis of the Manifestos in One Nation

We analyze *LPF* and *LN* as within-Netherlands Dutch-language parties. Distinguishing the two parties is a potentially difficult task, since they originally formed one party, the populist party *LN*, founded in June 2001, with Pim Fortuyn as party leader. After being ousted for blatant anti-Muslim comments, Fortuyn launched his own national party, *LPF*. While it is often claimed that *LN* is a *populist* rather than an extreme right party, *LPF* can be expected to be slightly more right-wing (Buyse, 2002). Yet, Fortuyn was openly homosexual and advocated liberal social values, which are very different from traditional right-wing values. In order to check if the AID methods could distinguish between the two manifestos, a subpart of the *LN*-manifesto was coded 'anonymous', while the other subparts (i.e. the other 9 *LN*-subparts and the *LPF*-part) were given an arbitrary code (i.e. l for *LN* and p for *LPF*). The task given to the program is to assign the subpart to *LN* instead of to *LPF*, using AID methods.

We concordanced the manifesto subparts using letter unigrams, letter bigrams and word unigrams. Then, both the *similarity ratio* and *Mann-Whitney* were calculated.

	Letter unigrams	Letter bigrams	Word unigrams
Ratio	1.299	1.131	1.028
Ranking	ln4 fits in category l ln4 fits in category p	ln4 fits in category l ln4 fits in category p	ln4 fits in category l ln4 fits in category p
Mann-Whitney	$p < 0.0005$	$p < 0.025$	$p < 0.25$

Table 1: Results of AID-tests classification of 'anonymous' subtext ln4 to *LN* (l) vs. *LPF* (p)

[4] In authorship attributions, the category corresponds to author identity.

[5] This is inspired by the proposal of Kilgarriff (1996) for equally-sized subcorpora.

[6] For the corpus analysis in one nation (§4.2) and in one language (§4.3), repeated tests for several subparts, for a communist party and for manifestos of Germany, Austria and Great-Britain gave similar results (Van Gijsel, 2002).

The ranking indicates that all three tests correctly attribute the subpart to the correct manifesto of *LN*, with a higher ratio measure for letter unigrams, followed by letter bigrams, indicating that letter unigrams perform best. Similarly, the output of *Mann-Whitney* shows that the attribution is highly significant for letter unigrams ($p < 0.0005$),[7] while letter bigrams are also significant ($p < 0.025$). By this test, a word unigram count is not significant ($p < 0.25$). These results cross-validate recent AID work, specifically McCombe's (2002) results. More importantly, the consistent correct attribution points at the *internal homogeneity* of the manifestos, which can therefore be considered fully-fledged corpora.

4.3 Analysis of the Manifestos in One Language

Since we intended to compare right-wing parties cross-nationally, in a second step, the manifestos in one *language* were analyzed similarly. Supplementing the manifestos of The Netherlands, the manifesto of the traditionally fascist Flemish party *VB* (*Flemish Block*) is analyzed. The attributional results for subparts of the *VB*-manifesto are consistent, validating it as a corpus. To further verify if the AID methods are robust enough to cope with the interference of inherently nation- and context-dependent elements in the manifestos, a communist manifesto, of the Flemish party *PvdA (Partij van de Arbeid/Labour Party)*, was included as a dummy. Although the attributional results for subparts of *PvdA* are not significant, repeated tests make clear that in the output ranking, the communist party is not closer related to the other Flemish party, *VB*, than to the Dutch parties, suggesting that a cross-national extension of the right-wing manifesto analysis is viable. Thus, internal homogeneity of the manifestos on a cross-national level shows the within-corpus differences to be smaller than between-corpora differences legitimating measurement of distances among the representative corpora, so that in a next step, the similarity among the corpora could be measured as an indication of the parties' political and ideological positions.

5 Placing Right-Wing Parties in a Left-Right Spectrum

In this section we discuss the distance between the parties as measured by treating the manifestos in their entirety as corpora. In general, statistical methods to reliably measure the distance between small, unequally sized corpora are scarce, Kilgarriff (2001) proposed χ^2 as a 'single measure' of distance between internally homogeneous corpora. The pairwise similarity ranking based on $\frac{\chi}{d.f.}$, is interpreted as indicating the level of similarity among the manifestos. Here, the manifestos in their entirety are compared, by letter unigrams, which consistently emerged as the clearest method to distinguish between them. Although the analysis would clearly benefit from a better similarity measure, enabling the direct statistical comparison of a number of corpora cross-linguistically, this measure will be interpreted as indicating the distances among the texts.

	$\frac{\chi}{d.f.}$	p-value
LN-VB	15.61	$p < 0.0001$
LPF-VB	8.18	$p < 0.001$
LN-LPF	2.97	$p > 0.05$

Table 2: Results of the inverse similarity ranking of the Dutch parties

The inverse similarity ranking shows that the difference between *LPF* and *LN* is not significant, on the basis of letter unigram frequencies. The difference between *LPF* and *VB* was significant at a 0.001 level and between *LN* and *VB* even at a 0.0001 level. These figures tie in

[7] Note that p measures the probability that the similarity judgement is due to mere chance.

with background knowledge: *LN* and *LPF* are both populist, 'new style' right wing parties, combining strong anti-immigration views with liberal social values, while *VB* is a traditional fascist party. Further, as was said before, *LPF* is more right-wing than *LN*, which is also clear from the higher similarity score for *LPF* with *VB*.[8]

Similar analysis was carried out for the manifestos translated in English (Van Gijsel, 2002, pp. 93-95),[9] enabling extension of the *cross-national* analysis. Again, the inverse similarity ranking brings out the difference between 'traditional' right-wing parties, like for example *BNP*, and the more populist, new-style right-wing parties, like *FPÖ*, which eclectically combine strong anti-immigration views with liberal social values; a difference which could not be taken into account with an *a priori* analysis trying to position the parties on pre-established policy dimensions.[10]

6 Conclusion

We have described our attempts to locate a number of European right-wing parties in single cline, analyzing their manifestos using tools of corpus linguistics. To verify applicability of corpus techniques, we applied AID methods to establish that intra-category differences are smaller than inter-category distances among the texts. This confirmed again that AID methods using letter frequencies are highly reliable, and verifies the internal homogeneity of the manifestos as corpora. The results, which show that the manifesto analysis as measured by $\frac{\chi}{d.f.}$differentiates 'traditional' and 'new-style' right-wing parties, demonstrate that a fully computerized analysis (specifically lacking content analysis) can give insight in the relatively unresearched policy space of right-wing parties. However, the analysis could benefit from methodological improvements and a cross-linguistic extension of the statistical measure. This work illustrates the use and limits of automated corpus linguistic techniques for small, unequally-sized 'real language' data sets.

References

Budge, I. (2001). Validating the MRG approach. In Laver, M. (Ed.), *Estimating the Policy Position of Political Actors*, pp. 3–9. London: Routledge/ECPR Studies in European Political Science.

Buyse, A. (Ed.). (2002). *Nieuw Radicaal Rechts in Europa*. Antwerpen/Amsterdam: Houtekiet.

Canter, D. (1992). An Evaluation of the "Cusum" Stylistic Analysis of Confessions. *Expert Evidence*, *1*(3), 93–99.

Chaski, C. (1998). A Daubert-Inspired Assessment of Current Techniques for Language-Based Author Identification. In *ILE Technical Report 1098*, pp. 97–148.

de Vries, M. (1999). *Governing with Your Closest Neighbour: An Assessment of Spatial Coalition Formation theories*. Ph.D. thesis, UB Nijmegen.

Farringdon, J. M. (1996). *Analysing for Authorship*. Cardiff: University of Wales Press. With contributions by Morton, A.Q., M.G. Farringdon and M.D. Baker.

Forsyth, R. S. (1997). Short Substrings as Document Discriminators: An Empirical Study. Paper presented at ACH-ALLC'97.

Foster, D. (2001). *Author Unknown. On the trail of Anonymous*. Macmillan: London, Basingstoke and Oxford.

Holmes, D. I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, *13*(3), 111–117.

Khmelev, D. & Tweedie, F. J. (2001). Using Markov Chains for Identification of Writers. *Literary and Linguistic Computing*, *16*(3), 299–308.

[8] Setting aside political content of 'left' or 'right' and taking $\sim$ to represent similarity; $<$, strict difference, we can make the following inference from pairwise comparisons: $LPF \sim LN, LPF < VB, LN \ll VB \models LN \preceq LPF$. Strikingly, the uninterpreted similarity partial ordering fits intuitions about orderings informed by political content.

[9] The manifestos of the English parties *NF* (National Front) and *BNP* (British National Party), of the Austrian *FPÖ* (Freedom Party), the French *FN* (Front National/National Front) and of *LPF* are not reliably discriminable.

[10] It can be remarked that cross-linguistically, only impressionistic conclusions are possible, linking a non-English manifesto such as for example the *LN*-text, which is known to be 'populist' and closely related to *LPF*, rather to *FPÖ*, which is closer to *LPF*, than to *BNP*.

Kilgarriff, A. & Salkie, R. (1996). Corpus Similarity and Homogeneity via Word Frequency. In *Proceedings of Euralex 96*.

Kilgarriff, A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches.. In *Language Engineering for Document Analysis and Recognition*. Proceedings, AISB Workshop, Falmer, Sussex.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, *6*(1), 97–133.

Laver, M. (Ed.). (2001a). *Estimating the Policy Position of Political Actors*. Routledge.

Laver, M. (2001b). Position and Salience in the Policies of Political Actors. In Laver, M. (Ed.), *Estimating the Policy Position of Political Actors*, pp. 66–75. London: Routledge/ECPR Studies in European Political Science.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, *97*.

Laver, M. & Garry, J. (2000). Estimating Policy Positions from Political Texts. *American Journal of Political Science*, *44*(3), 619–634.

Mair, P. (2001). Searching for Positions of Political Actors. In Laver, M. (Ed.), *Estimating the Policy Position of Political Actors*, pp. 33–49. London; Routledge/ECPR Studies in European Political Science.

McCombe, N. (2002). Methods of Author Identification. B.A. (Mod) CSLL Final Year Project, TCD.

Morton, A. Q. (1986). Once. A Test of Authorship Based on Words Which Are Not repeated in the Sample. *Literary and Linguistic Computing*, *1*(1), 1–8.

Mosteller, F. & Wallace, D. (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Reading: Addison-Wesley.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press.

Van Gijsel, S. (2002). A Corpus Linguistic Analysis of European Right-Wing Party Manifestos. Master's thesis, Centre for Language and Communication Studies, Trinity College, University of Dublin.

Yule, G. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.